DOCUMENT RESUME

ED 384 653 TM 023 886

AUTHOR Zwick, Rebecca
TITLE Stability of DIF Classification: An Alternative Representation of the Variability of the Mantel-Haenszel DIF Statistic.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 19 Apr 95
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Adaptive Testing; *Bayesian Statistics; *Classification; Computer Assisted Testing; Estimation (Mathematics); *Item Bias; Probability; Sample Size; *Sampling; Simulation; *Statistical Distributions
IDENTIFIERS Educational Testing Service; *Mantel Haenszel Procedure; *Variability

ABSTRACT

This paper describes a study, now in progress, of new methods for representing the sampling variability of Mantel-Haenszel differential item functioning (DIF) results, based on the system for categorizing the severity of DIF that is now in place at the Educational Testing Service. The methods, which involve a Bayesian elaboration of procedures developed by Zwick, Thayer, and Wingersky (1994) are expected to yield more interpretable results in small samples than do conventional procedures. They may therefore be useful in the small-sample conditions that occur in adaptive testing. Two methods are suggested. One is for obtaining a probabilistic estimate of the true DIF category, and the other is for predicting the future observed category. These Bayesian methods have not yet been tested through simulation. Some potential limitations are discussed, subject to testing the accuracy of the probabilistic results and the sensitivity of results to the shape and parameters of the prior distribution. Two tables and three figures illustrate this exploration of techniques. (Contains 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

March 14, 1995

Stability of DIF Classification: An Alternative Representation of the Variability of the Mantel-Haenszel DIF Statistic

Rebecca Zwick

Educational Testing Service

ED 384 653

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

R. Halpern

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Presented at the annual convention of the National Council on Measurement in Education, San Francisco, April 19, 1995. I am thankful to Charlie Lewis for his valuable contributions to the Bayesian approach described here, to Dorothy T. Thayer for providing statistical programming, creating graphical displays, and reviewing the manuscript, and to the Graduate Record Examinations program for making DIF results available. Requests for copies of this paper should be sent to Rebecca Zwick, ETS, Princeton, NJ 08541 or via Internet to rzwick@ets.org.

TM023886

Stability of DIF Classification: An Alternative Representation of the Variability of the Mantel-Haenszel DIF Statistic

This paper describes a study, now in progress, of new methods for representing the sampling variability of Mantel-Haenszel differential item functioning (DIF) results, based on the system for categorizing the severity of DIF that is now in place at Educational Testing Service (ETS). The methods, which involve a Bayesian elaboration of procedures developed by Zwick, Thayer, and Wingersky (1994a; 1994b; in press), are expected to yield more interpretable results in small samples than do conventional procedures and therefore may be advantageous under the small-sample conditions that occur in adaptive testing. More generally, the methods may convey information about the sampling variability of DIF results in a more comprehensible way than do the current procedures.

The first section of the paper gives a brief description of the Mantel-Haenszel DIF procedure and the existing ETS system for classifying the degree of DIF. In order to ground the subsequent discussion, the second section gives an illustration of the results of applying one of the proposed Bayesian procedures and provides a rationale for the approach. The third section describes the development of the Bayesian methods and presents the underlying statistical model. The final section lists a number of unresolved issues and possible drawbacks of the proposed approach.

The Mantel-Haenszel Method of DIF analysis

The Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) is a well-established method for assessing DIF. A $2 \times 2 \times K$ table of examinee data is constructed based on item performance (right or wrong), group membership (the

focal group, which is of primary interest, or the *reference group*), and score on an overall proficiency measure (with K levels). The MH (1959) odds ratio estimate is then used to compare the two groups in terms of their odds of answering the item correctly, conditional on the proficiency measure. The MH index of DIF, *MH D-DIF*, is proportional to the natural log of the MH odds ratio estimate $\hat{\alpha}_{MH}$; the transformation of $\hat{\alpha}_{MH}$ places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). *MH D-DIF* is defined so as to be negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. Phillips and Holland (1987) derived an estimated standard error for $\ln(\hat{\alpha}_{MH})$; their result proved to be identical to that of Robins, Breslow and Greenland (1986).

Classification of DIF Severity

ETS has a system for categorizing the severity of DIF based on both the *magnitude* of the *MH D-DIF* index and the *statistical significance* of the results (see Zieky, 1993). According to this classification scheme, a "C" categorization, which represents moderate to large DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = .05$). A "B" categorization, which indicates slight to moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = .05$) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to be free of DIF. Items that fall in the C category are typically eliminated from tests or subjected to further scrutiny.

Two Bayesian variants of this classification system will be presented: Method 1, the simpler of the two, provides a probabilistic estimate of the item's *true* DIF category. This procedure uses only the magnitude criteria above. Method

2 uses both the magnitude and statistical significance criteria to predict the item's future *observed* category in probabilistic terms. For both methods, the Bayesian calculations are based on the observed *MH D-DIF* statistics and their standard errors.

Illustration: Results of Applying Method 1

Before explaining the statistical underpinnings of the Bayesian procedures, an illustration of the results of applying Method 1 to actual test data will be presented. The method was applied to MH DIF results for the 76 items in the verbal section of the 1990 Graduate Record Examinations (GRE); Table 1 shows the results for four items. The first three columns give an item number, the originally obtained *MH D-DIF* statistic and the original DIF category, respectively. The next three columns give the Method 1 results: a probabilistic estimate of the item's true DIF category, expressed in the percent metric. These percents are based on an estimate of the distribution of the true (and unknown) MH DIF parameter and on the magnitude criteria for DIF classification. (Discussion of the Method 2 results, which appear in the next three columns, will be deferred until a later section.)

Two important aspects of the probabilistic results are the true DIF category that is most likely (indicated by the underlined percents) and the degree of concentration of the results. First consider Items 1 and 2, which were designated as A items using the ordinary procedure. The *MH D-DIF* values for these items, which are very similar, are only slightly below the criterion value of 1 needed to achieve B status. Although the Bayesian procedure leads to the conclusion that the most likely true category for both items is A, the degree of certainty is smaller for the first item (80%) than for the second item (91%). The difference is even more dramatic for the two C items, 3 and 4. Although Item 3 has a smaller *MH D-DIF*

value than Item 4, the true category is much more likely to be C for Item 3 than for Item 4 (85% versus 6%). In fact, the results tell us that the true category for Item 4 is most likely to be A.

The tabled results show a pair of A items and a pair of C items. Within each pair, the original DIF results were similar, but the Bayesian results are different, particularly for the C items. This disparity is a result of the differences in the sample sizes on which the original DIF findings were based. For Items 1 and 3, the displayed DIF analyses compared men (n = 4,253) and women (n = 6,491); the values of *SE (MH D-DIF)* were .20 and .16, respectively. The DIF results for Items 2 and 4 were based on analyses of White (n = 8,736) and Asian (n = 220) examinees; *SE (MH D-DIF)* values were .39 and .55, respectively. The precise role of *SE (MH D-DIF)* in the Bayesian analysis is detailed in a later section.

Rationale for the Proposed Bayesian DIF Methods

Why should we consider substituting probabilistic results like those of Table 1 for the currently reported DIF results, which include *MH D-DIF*, *SE (MH D-DIF)*, and the DIF category? For some consumers of DIF results, such as test developers and clients, the proposed approach may convey information about DIF stability in a more useful way. The results make it clear that an item's DIF category is not known with certainty: The conclusion about the item's true DIF status is probabilistic, not deterministic. Some items have concentrated results, indicating a high degree of certainly, while other items have more diffuse results. DIF categories can be presented graphically as in Figure 1; such displays may provide a more intuitive representation of the degree of uncertainty associated with item classification. The goal of drawing more attention to the variability of DIF findings could, of course, be achieved by providing confidence intervals for the *MH D-DIF* statistic rather than by expressing the variability in terms of the A,

B, and C categories. However, the current ETS DIF policy, which determines whether items should be scrutinized by committees or even removed from the test, is based on the classification results: It is the A, B, and C designations that form the basis of decision-making.

The proposed methods do not depend on the *particular* rules now used to assign items to categories; they can be modified easily if the current criteria are adjusted or if other hypothesis-testing approaches are substituted for the MH procedure. In fact, application of these Bayesian methods may provide "diagnostic" information on properties of the rules: We may question a system that classifies Item 4 in Table 1 as a C if, under some reasonable model, its most likely true category is A.

Development of the Bayesian Methods

The precursor to the Bayesian methods proposed here was developed in a simulation study of methods for assessing DIF in computer-adaptive tests (Zwick, Thayer, & Wingersky, 1994a; 1994b; in press). A brief description of this procedure is given here; the details and justification can be found in the original references. In the simulation study, we "administered" a computer-adaptive test (CAT) to 60,000 simulated reference group examinees and 60,000 examinees from each of several focal groups. Because of the adaptive nature of the test, the sample sizes per item (within population group) had an extremely large range. Because our goal was to investigate the behavior of DIF statistics for certain target sample sizes, it would not have been useful simply to analyze the available data for each item. We therefore adopted the *expected table* (ET) method, an approach that allowed us to project the behavior of the DIF statistics for our target sample sizes.[1]

[1]The ET approach and the subsequent estimation of expected percents of A, B, and C results were suggested by C. Lewis.

The ET method (which is not a part of the proposed Bayesian approach) involved the following steps: (a) Responses to the CAT items were generated for each examinee. (b) For each of the items in the CAT pool, all the available data (out of a maximum of 60,000 responses per group) were used to form the $2 \times 2 \times K$ contingency table needed for DIF analysis. (c) The table frequencies were converted to proportions of the total number of observations for the group in question. (d) Using these proportions as estimates of the corresponding population probabilities (for the relevant configuration of simulation conditions), expected tables were obtained by multiplying the probability estimates for focal group cells by the target focal group sample size and then doing the same for the reference group cells. (e) ET DIF statistics and their standard errors (say, *MH D-DIF** and *SE*(MH D-DIF)*) were computed, based on the expected tables.

After the ET analyses, we used the *MH D-DIF** and *SE*(MH D-DIF)* values to calculate the expected percents of A, B, and C results for each item. We assumed that the distribution of the *MH D-DIF* statistic for this particular configuration of simulation conditions and target sample sizes was normal with mean *MH D-DIF** and standard deviation *SE*(MH D-DIF)*. We applied to this distribution a classification probability algorithm (CPA) that we had developed, which implements, in essence, the ETS rules for DIF classification and yields the expected percents of A, B, and C results for the items. Two supplementary studies of the CPA showed that it operated as intended.

Modification of the ET/CPA Approach for Application to Real Data

The success of the ET/CPA approach in the simulation study was a result of the large amount of data that was available for each item--typically at least 10,000 observations. This allowed us to treat *MH D-DIF** and *SE*(MH D-DIF)* as population parameters. Because most DIF analyses are based on much smaller

samples, a modification of the approach is needed to produce accurate estimates of the percents associated with the A, B, and C categories. Specifically, Bayesian priors are used to improve the estimates of the DIF distributions. The priors express our informed beliefs about the distribution of the DIF *parameters*. Combining the prior information with the observed DIF results allows us to derive estimates that are more stable than those based on the obtained DIF statistics alone.

A Bayesian approach to the categorization of DIF severity was discussed in 1987 in a series of ETS internal memos by Neil Dorans, Paul Holland, Gary Marco, and others, but was not implemented. An elegant empirical Bayes model for DIF estimation that did not involve the DIF categories was later developed and applied by Longford, Holland, and Thayer (1990; see also Longford, 1994). The main purposes of that work were to assess test-level DIF and to determine how much DIF variation was attributable to each of several sources, such as item difficulty and administration conditions. The models were necessarily quite elaborate, and the computational procedures involved estimation of the parameters of the prior distribution from the data, using an iterative procedure. By contrast, the model and computational procedures proposed here are intended to be very simple, so that the approach can be applied easily by testing programs. The prior mean and variance are treated as fixed, and the Bayesian estimates are obtained from the observed DIF statistics through a noniterative procedure. The Bayesian estimates are used primarily as a step toward the estimation of the probabilities associated with each of the categories of DIF severity.

In developing the proposed Bayesian version of A, B, and C classification, a question arose as to which DIF distribution was of primary interest in applications the analysis of actual test data. The goal of applying the ET/CPA approach in the simulation study was to predict the behavior of CAT DIF *statistics* for certain conditions and sample sizes. In a real-data applications, it may be of greater

interest to estimate the distribution of the *true* DIF parameters. This is the goal of Method 1. For situations in which the goal is indeed to predict the behavior of future *MH D-DIF statistics*, Method 2 can be applied. Although Method 2 is more analogous to the ET/CPA approach described above, Method 1 is presented first because of its greater simplicity.

Method 1: Obtaining a probabilistic estimate of the true DIF category. Method 1 uses the observed values of *MH D-DIF* and *SE (MH D-DIF)*, along with an assumed prior distribution, to obtain the posterior distribution of true MH DIF parameters. The magnitude criteria for assigning items to the A, B, and C categories, described above, are then applied to this posterior distribution. The statistical significance criteria are not relevant because it is the distribution of parameter values, not statistics, that is being estimated. The resulting A, B, and C percents can be regarded as representing our state of knowledge about the true DIF category for the item.

The model can be expressed as follows. (The notation change from *MH D-DIF* to *MH* makes the presentation less cumbersome.) We know that $\ln(\hat{\alpha}_{MH})$ has an asymptotic normal distribution (Agresti, 1990). Therefore, it is reasonable to assume that

$$MH_i|\theta_i \sim N(\theta_i, \sigma_i^2), \qquad (1)$$

where MH_i is the MH statistic for item i, $\sigma_i^2 \equiv SE^2(MH_i)$ is the squared standard error of the MH statistic, and $E(MH_i) = \theta_i$ represents the unknown parameter value corresponding to MH_i. We also assume the following prior distribution for θ_i :

$$\theta_i \sim N(\mu, \tau^2), \tag{2}$$

where μ is the across-item mean of MH statistics for an administration and $\tau^2 \equiv \text{Var}(MH)$ is the across-item variance. When MH DIF analyses are conducted using number-right score as the matching variable, the MH DIF statistics are constrained to sum to approximately zero over the set of items. Therefore, in many applications, setting μ equal to zero is appropriate and, for the present, we assume this to be the case. The selection of an appropriate value for τ^2 will be discussed below.

The posterior distribution of θ_i, given the observed statistic, MH_i, can now be expressed as

$$f(\theta_i \mid MH_i) \propto f(MH_i \mid \theta_i)\, f(\theta_i). \tag{3}$$

Standard Bayesian calculations (see, e.g., Novick & Jackson, 1974) show that this distribution is normal with mean $W_i MH_i$ and variance $W_i SE^2(MH_i)$, where

$$W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2} \equiv \frac{Var(MH)}{SE^2(MH_i) + Var(MH)}. \tag{4}$$

The means and variances for the model, prior, and posterior distributions, assuming $\mu = 0$, are summarized in Table 2. (The posterior predictive distribution, discussed below, is also included in the table.)

The posterior mean is a simple shrinkage estimator of Mantel-Haenszel DIF, requiring no iteration. As $SE(MH_i)$ approaches 0, W_i approaches 1 and the posterior mean approaches the observed value. On the other hand, the larger the

value of *SE* (MH_i), the more the Bayesian estimate is shrunk toward the prior mean of 0. Because the weights W_i vary across items, the rank order of the Bayesian estimates can differ from the order of observed MH statistics under this model. Figures 2 and 3 show the Bayesian estimates plotted against the observed DIF statistics for the Male-Female and White-Asian DIF analyses, respectively. Whereas the plot for the Male-Female analysis (Figure 2) is close to the 45-degree line, the plot based on the less stable White-Asian DIF results (Figure 3) clearly illustrates the shrinkage phenomenon.

Method 2: Predicting the Future Observed Category. Method 2 uses the observed values of *MH D-DIF* and *SE (MH D-DIF)*, along with an assumed prior distribution, to obtain the posterior *predictive* distribution of future values of the *MH D-DIF* statistic. Then, a version of the CPA is applied which appropriately incorporates both the magnitude criteria and the statistical significance criteria for assigning items to the A, B, and C categories. The resulting percents represent a prediction of the results of future DIF analyses. For example, Method 2 could be used to predict the number of A, B, and C items in a future test administration.

The model presented in equations 1 to 3 is assumed to hold. The posterior predictive distribution (see Novick & Jackson, 1974) of future MH_i values can then be obtained as

$$f(MH_i') = \int f(MH_i'|\theta_i) f(\theta_i|MH_i) d\theta_i.$$

Because of changes in sample sizes or other conditions, $f(MH_i'|\theta_i)$ may not be the same as $f(MH_i|\theta_i)$. For the present, we assume for simplicity that $f(MH_i'|\theta_i) = f(MH_i|\theta_i)$. Departures from this assumption will be discussed in the final section of the paper. Under the further assumption that $\mu = 0$, the mean and variance of

the posterior predictive distribution are given as $W_i MH_i$ and $(1+W_i)SE^2(MH_i)$, respectively, where W_i is defined as in equation 4 (see Table 2). The posterior predictive mean is identical to the posterior mean, but the posterior predictive variance is larger than the posterior variance. The results of applying Method 2 are given in the three rightmost columns of Table 1. Except in the case of Item 4, the Method 2 results are considerably more diffuse than the Method 1 results.

Computations for Methods 1 and 2. As already described, the means and variances for the posterior and posterior predictive distributions can be obtained from the formulas of Table 2. The only remaining aspect of the procedure that needs further explanation is the selection of a value for the prior variance τ^2. It is anticipated that in actual applications of these methods to test data, the results of previous DIF analyses will be used to inform the selection of an appropriate value. In the present example, the DIF results from the Male-Female and White-Asian analyses themselves were used to determine a reasonable value for τ^2. A simple relationship of τ^2 to the variance of the observed MH_i values was assumed: τ^2 = Var (MH_i) - Average (SE_i^2). That is, the across-item prior variance of the true DIF values, τ^2, is obtained by deflating the across-item variance of the observed DIF statistics, Var (MH_i), by an amount equal to the average sampling variance, Average (SE_i^2). Applying this formulation to the Male-Female and White-Asian DIF results led to the conclusion that a prior variance of τ^2 = .25 would be reasonable in both cases. This value is similar to that obtained by Longford (1994), who used an iterative procedure to estimate τ^2 for these data sets. One way to understand the strength of this prior is to note that, if the true DIF distribution is normal with mean 0 and variance .25, then about 5 % of the values are greater than 1 in absolute magnitude and about 1/4 % of the values are greater than 1.5 in absolute magnitude.

Table 1 shows the values of W_i, $W_i MH_i$, $\sqrt{W_i}SE_i$, and $\sqrt{(1+W_i)}SE_i$ for each of the four exemplar items. As an illustration of the required computations, consider Item 3. The observed statistics are $MH_3 = -1.82$ and $SE_3 = .16$. Application of the formulas in Table 2 shows that $W_3 = .91$ and the posterior mean and standard deviation are $W_3 MH_3 = -1.66$, and $\sqrt{W_3}SE_3 = .15$, respectively. In a normal distribution with this mean and standard deviation, the percent of values with absolute magnitude less than 1 (A) is nearly 0, the percent with absolute magnitude between 1 and 1.5 (B) is about 15, and the percent with absolute magnitude greater than 1.5 (C) is about 85. These are the percents given in the Method 1 results in Table 1.

The posterior predictive mean is the same as the posterior mean; the posterior predictive standard deviation is $\sqrt{(1+W_3)}SE_3 = .22$. Application of the CPA to a normal distribution with these parameters leads to predicted percents of 0, 23, and 76 in the A, B, and C categories, respectively. These are the Method 2 results in Table 1.

Note that, contrary to intuition, larger sample sizes need not lead to greater certainty of classification. For example, the Table 1 results indicate that Item 2 is more certainly an A than Item 1, despite the smaller sample size for Item 2. This phenomenon can occur because the certainty of classification depends not only on the variability of the DIF distribution, but on the proximity of the posterior (predictive) mean to the cut-offs for item classification. In this case, the posterior mean for Item 2 is farther from the value of 1 needed for B classification than the posterior mean for Item 1.

Cautions and Unresolved Issues

It must be noted that this research has just begun and the Bayesian methods, unlike their precursor, have not yet been tested via simulation. Future work will

involve testing the accuracy of the probabilistic results and the sensitivity of results to the shape and parameters of the prior distribution.

A possible drawback of the Bayesian approaches is that they rely on the asymptotic normality of MH statistic, which may impair their performance in small samples--the very circumstance in which they are potentially most useful. This too can be studied through simulation. The methods also ignore the error associated with the estimation of *SE(MH D-DIF)*. However, Longford (1994) investigated the impact of treating *SE(MH D-DIF)* as known in the Bayesian approach of Longford, Holland, and Thayer (1993) and concluded that it was minimal.

A limitation of Method 2 as presented here was the assumption that the distribution of a future *MH D-DIF* statistic for a given item would be the same as the distribution of the *MH D-DIF* statistic that had already been observed. Clearly, this assumption would be violated in many cases. In particular, if sample sizes changed between one administration and the next, the assumption would not hold. However, with only minor modification to the algebra, the method can accommodate the case in which the future standard error is expected to differ from the observed standard error. Apart from sample size, the mode of test administration can also affect *SE(MH D-DIF)*. For a given sample size and set of item parameters, *SE(MH D-DIF)* tends to be smaller in adaptive than in nonadaptive administration (Way, 1994; Zwick, 1994; Zwick, Thayer, & Wingersky, 1994b). This would be relevant if DIF results from a paper-and-pencil administration of a test were to be used to predict DIF results in a CAT version of the test.

Modification of the proposed methods will be needed to allow for application of these approaches when an external test is used as a matching variable (and hence μ cannot be assumed to be equal to zero), and when there is variation across test administrations in item context or examinee populations.

These factors would come into play, for example, if DIF results from a pretest administration were to be used to predict DIF results in an operational administration.

Another worthwhile elaboration of the Bayesian methods would be the incorporation of system for modifying the prior distribution each time an item was given. This would provide a means of cumulating DIF information across all administrations of an item and would take fuller advantage of the capabilities of Bayesian techniques.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty.* ETS Research Report No. 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Longford, N. T. (1994). *A variance measure of test-level differential item functioning.* Unpublished manuscript.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning*, p. 171-196. Hillsdale, NJ: Erlbaum.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425-431.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*, 311-323.

Way, W. D. (1994). *A simulation study of the Mantel-Haenszel procedure for detecting DIF with the NCLEX using CAT.* Technical report, Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland and H. Wainer (eds.), *Differential item functioning*, pp. 337-347. Hillsdale, NJ: Erlbaum.

Zwick, R. (1994). *The effect of the probability of correct response on the variability of measures of differential item functioning.* ETS RR 94-44. Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994a). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121-140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994b). *DIF analysis for pretest items in computer-adaptive testing.* ETS Research Report No. 94-33. Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Wingersky, M. (in press). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement.*

Table 1

Result of Application of Bayesian Methods to DIF Results for Verbal Items from the 1990 Graduate Record Examinations

Item	Observed Results			Quantities Needed for Bayesian Classification				Method 1: Estimate of True Category (%)			Method 2: Prediction of Future Observed Category (%)		
	MH D-DIF	*SE(MH D-DIF)*	Category	W_i	$W_i MH_i$	$\sqrt{W_i}SE_i$	$\sqrt{(1+W_i)}SE_i$	A	B	C	A	B	C
1	0.98	0.20	A	0.86	0.84	0.19	0.28	80	20	0	72	27	1
2	0.94	0.39	A	0.62	0.58	0.31	0.50	91	9	0	80	19	2
3	-1.82	0.16	C	0.91	-1.66	0.15	0.22	0	15	85	0	23	76
4	-2.00	0.55	C	0.46	-0.91	0.37	0.66	60	35	6	60	34	7

*Because of rounding, the A, B, and C percents do not always sum to 100. Results for Items 1 and 3 are based on DIF analyses comparing men (n = 4,253) and women (n = 6,491). Results for Items 2 and 4 are based on analyses of White (n = 8,736) and Asian (n = 220) examinees.

Table 2

Means and Variances of Key Distributions

Distribution	Mean	Variance
Model $f(MH_i \mid \theta_i)$	θ_i (unknown)	$\sigma_i^2 = SE^2(MH_i)$
Prior $f(\theta_i)$	$\mu = 0$	$\tau^2 = Var(MH)$
Posterior $f(\theta_i \mid MH_i)$	$\frac{\tau^2}{\sigma_i^2+\tau^2} MH_i + \frac{\sigma_i^2}{\sigma_i^2+\tau^2}\mu$ $= \frac{Var(MH)}{SE^2(MH_i)+Var(MH)}(MH_i)$ $\equiv W_i MH_i$	$\frac{\sigma_i^2\tau^2}{\sigma_i^2+\tau^2} = \frac{Var(MH)\cdot SE^2(MH_i)}{Var(MH)+SE^2(MH_i)}$ $= W_i SE^2(MH_i)$
Posterior Predictive $f(MH_i')$	$\frac{\tau^2}{\sigma^2+\tau^2} MH_i + \frac{\sigma^2}{\sigma^2+\tau^2}\mu$ $= W_i MH_i$	$\frac{\sigma^2\tau^2}{\sigma^2+\tau^2} + \sigma^2 = \sigma^2\left(1+\frac{\tau^2}{\sigma^2+\tau^2}\right)$ $= (1+W_i)SE^2(MH_i)$

Figure 1

Method 1 Results

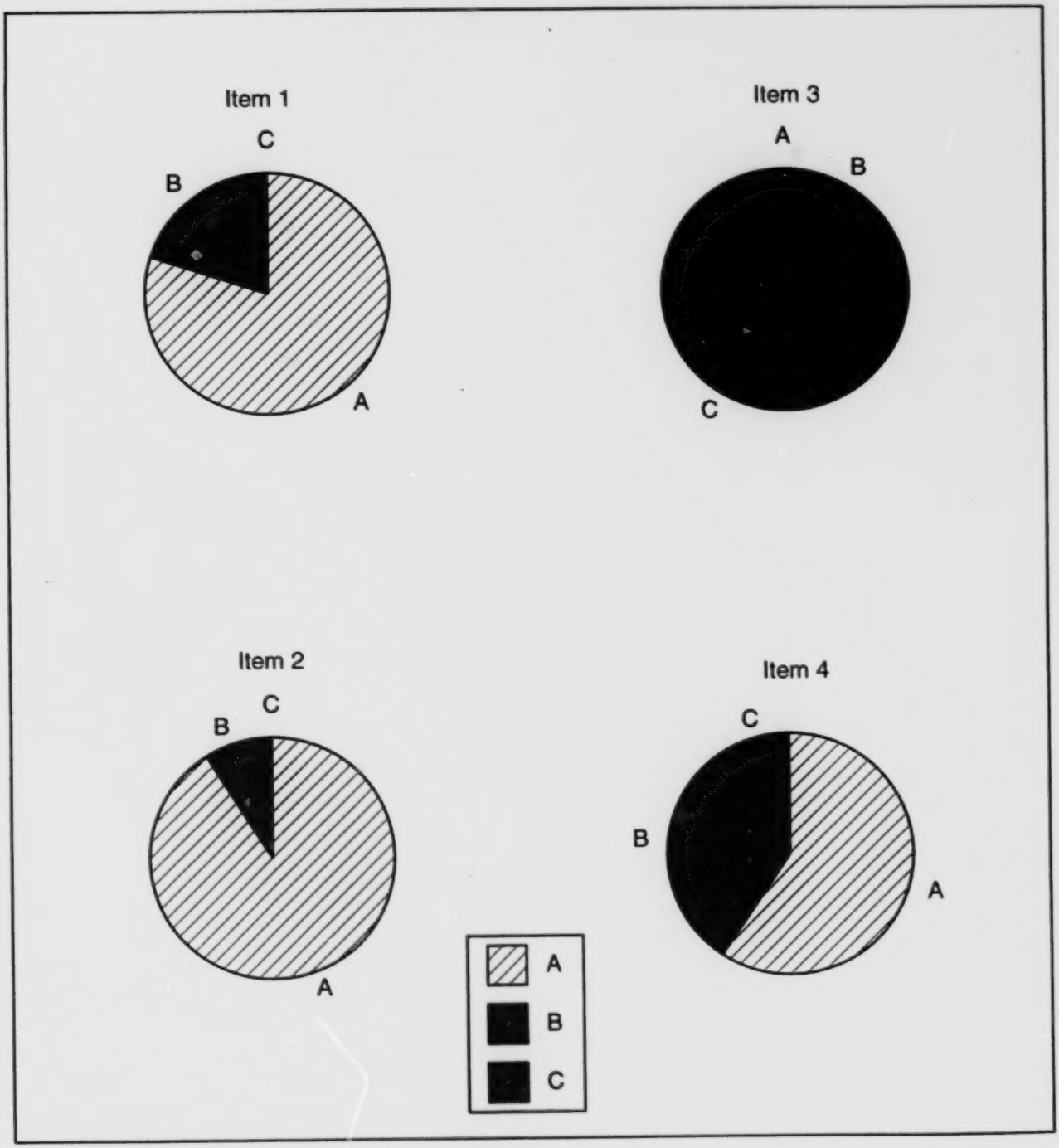


Figure 2

MH AND ADJUSTED MH FOR MALE-FEMALE ANAYSIS OF GRE-V

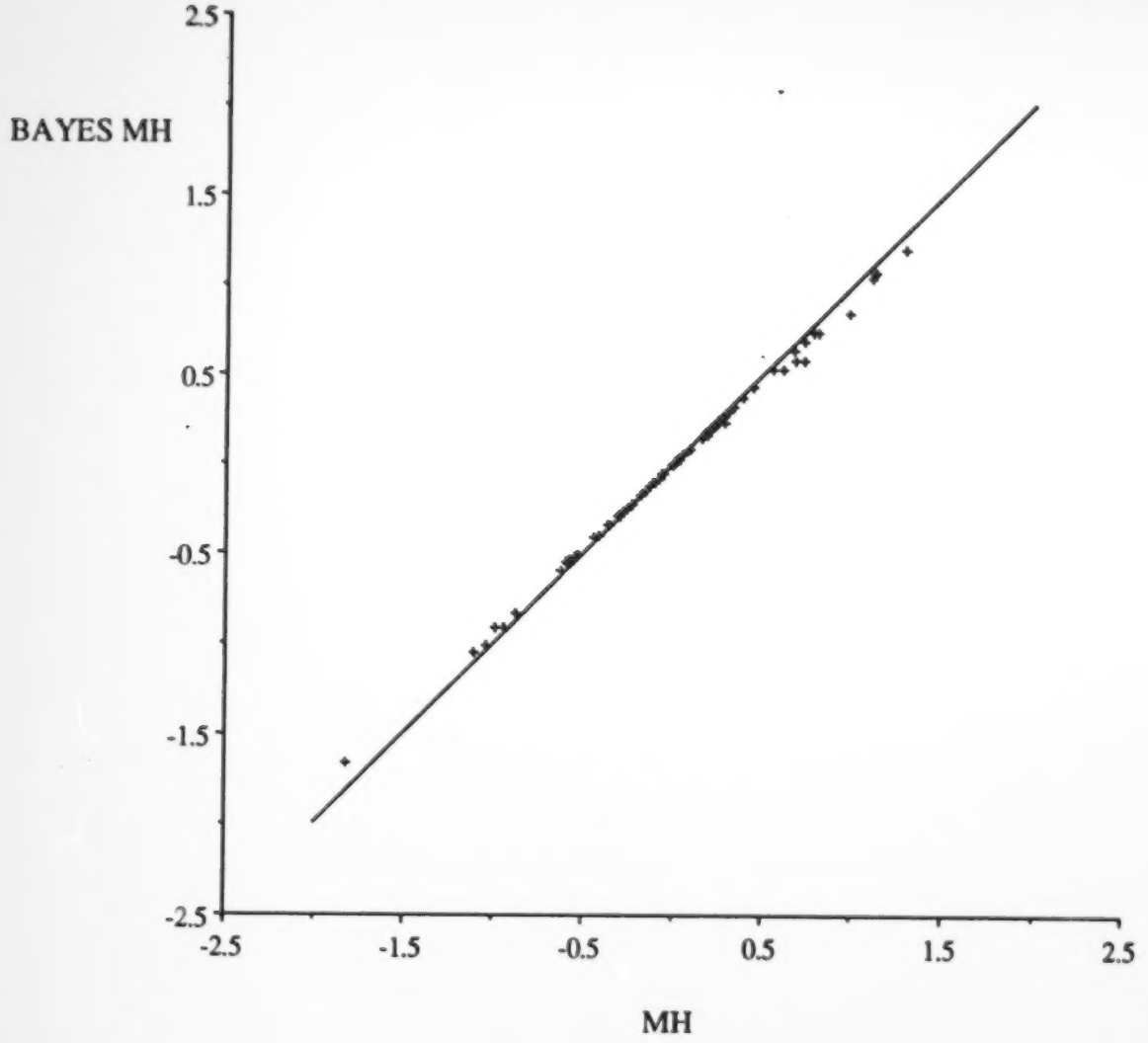
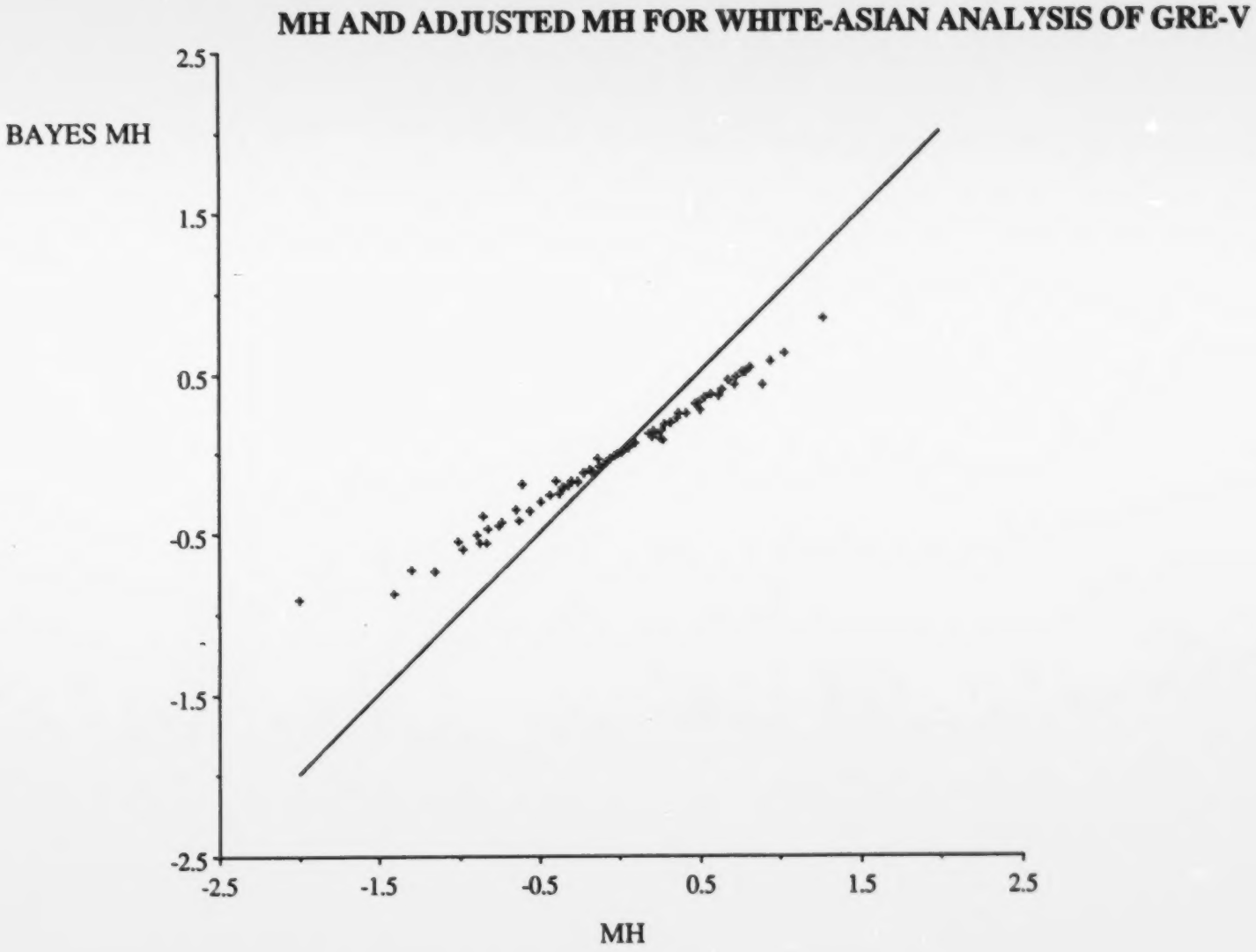


Figure 3


MH AND ADJUSTED MH FOR WHITE-ASIAN ANALYSIS OF GRE-V
BAYES MH
2.5
1.5
0.5
-0.5
-1.5
-2.5
-2.5
-1.5
-0.5
0.5
1.5
2.5
MH